Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle

Aicha SOUDANI[1,2] Aicha.soudani@hotmail.fr
Yosra MEHERZI[1,2] Yosra.meherzi@gmail.com
Asma BOUHAFS[1] Asma_bouhafs@yahoo.com
Francesca FRONTINI[3] Francesca.frontini@univ-montp3.fr
Carmen BRANDO[4] Carmen.brando@ehess.fr
Yoann Dupont[2] Yoa.dupont@gmail.com
Frédérique Mélanie-Becquet[2] Frederique.melanie@ens.fr

(1) *ECSTRA, IHEC, Université de Carthage*
(2) *Lattice UMR 8094, Université Paris 3-Sorbonne Nouvelle*
(3) *Praxiling UMR 5267, Université Paul-Valéry de Montpellier*
(4) *CRH UMR 8558 / Plateforme géomatique, EHESS*

Résumé

Dans cet article, nous proposons une chaîne de traitement reposant sur deux outils existants, l'un pour la reconnaissance des entités nommées, et l'autre pour la résolution des entités nommées. Par la suite, l'évaluation et l'adaptation de ces systèmes à l'analyse des textes issus de la littérature française du 19ème siècle sont présentés. Le résultat fourni par notre chaîne de traitement propose une visualisation projetant les entités nommées de type Lieu sur une carte et nous montrons enfin l'intérêt de ce travail pour les humanités numériques.

Abstract

In this article, we propose a processing pipeline relying on two existing tools, one for named-entity recognition, and the other for named-entity linking. We first present the evaluation and the adaptation of these systems to the analysis of texts from 19th Century French literature. We then show the result provided by the pipeline, namely a projection of the place entities onto a map. Finally, we discuss the interest of this work for the digital humanities.

Mots-clés : reconnaissance des entités nommées, résolution des entités nommées, cartographie.
Keywords : entity recognition, named entity linking, cartography.

Introduction

Les outils du traitement automatique du langage naturel (TAL) occupent une place importante dans le spectre des humanités numériques contribuant notamment à l'analyse d'œuvres littéraires numérisés sous l'angle des entités nommées (EN) [Lecluze et al, 2014]. Dans ce papier, nous nous intéressons particulièrement à la reconnaissance des entités nommés (REN) et la résolution des entités nommées (NEL), tâches répandues en TAL, afin d'enrichir des textes issus du canon littéraire français du 19ème siècle. D'une part, la REN consiste à identifier et catégoriser des expressions linguistiques comme les noms de personne, de lieu, et d'institution, d'autre part, la NEL vise à déterminer l'identité des entités, mentionnées dans le texte, à partir d'une base de connaissances (BC) telle que DBpedia, Wikidata, Yago, Geonames.

Dans ce travail, nous proposons une chaîne de traitement reposant sur deux outils REN et NEL qui seront adaptés à l'analyse de textes de la littérature française. Nous avons constitué un *gold standard* composé de deux chapitres du roman « Le ventre de Paris » d'Emile Zola et le premier chapitre du roman « César Birotteau » d'Honoré de Balzac. Nous avons également développé une application en ligne afin de proposer un rendu dynamique projetant les lieux repérés dans les textes sur une carte. Le restant de ce papier se décompose en trois parties. Les deux premières parties se consacrent à la présentation des outils REN et NEL sélectionnés pour notre

étude, et la dernière partie présente une conclusion évoquant l'intérêt de ce travail pour des projets en humanités numériques en cours.

1. Chaîne de traitement proposée

La chaîne de traitement proposée est illustrée en Figure 1, elle intègre le système « SEM » [Dupont 2017] pour la tâche REN et « REDEN » [Frontini et al 2015 ; Brando et al 2016] pour la tâche NEL. Le choix du format XML/TEI, incontournable pour l'édition numérique de textes, est le choix d'encodage fait pour l'entrée de notre chaîne. Il est donc indispensable que les deux outils supportent ce format. Pour REDEN, la question ne se pose pas. Par contre, il a fallu adapter SEM afin de donner support au format XML/TEI.

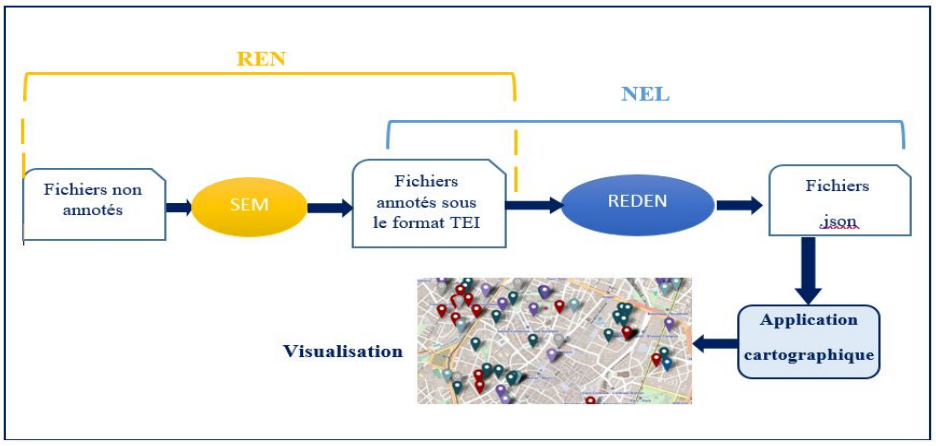

Figure 1. Chaîne de traitement proposée

1.1. Adaptation et évaluation du système SEM

SEM est un système de REN qui s'appuie sur une approche d'apprentissage supervisée [Raymond et al., 2010]. L'apprentissage du système repose sur l'entraînement d'un modèle à partir d'exemples afin de reproduire une tâche de prédiction. Un modèle pour SEM est disponible, entraîné à partir de textes journalistiques du French TreeBank [Sagot et al., 2012]. Pour ce travail, nous avons entraîné un modèle REN pour SEM à partir des textes littéraires français que nous avons manuellement annotés (le *gold standard*) par deux annotateurs distincts.

L'accord inter-annotateur donne les valeurs suivantes: 0,88 pour le rappel, 0,96 pour la précision et 0,91 pour la F-mesure. La proximité des valeurs de l'accord inter-annotateur, tous proches de 1, déduit la similarité de l'annotation produite par les deux experts.

Nous avons évalué les performances de ce modèle en termes de rappel, de précision et de F-mesure, ainsi qu'analysé les erreurs d'annotation récurrentes. Les expérimentations sur SEM concernent deux types d'entités nommées, à savoir personnes et lieux, et sont effectués sur le *gold standard.* Elles se divisent en trois grandes parties, décrites ci-dessous.

(1) Évaluation de SEM avec le modèle French Tree Bank :

Cette expérience montre que le modèle entraîné sur des textes journalistiques contemporains n'est pas suffisamment portable sur des textes littéraires (cf. tab 1). En particulier, les EN de type « Person » pose plus de problèmes avec des résultats de F-mesure variables qui peuvent atteindre un minimum de 10%. Cela vient du fait que les noms de personnes correspondent à des noms fictifs. Cependant, pour la reconnaissance des lieux, les résultats sont meilleurs avec des valeurs de F-mesure qui varient entre 24% et 33%. Ceci est dû au fait que les noms des lieux représentent des lieux réels existants et donc connus et appris par le modèle et présents dans le dictionnaire. Le tableau 1 suivant montre les résultats des expériences.

Tableau 1. Résultats expérimentations SEM

	SEM modèle FrenchTreeBank : Zola		SEM modèle FrenchTreeBank : Balzac		Adaptation 1	Adaptation 2
	Location	Person	Location	Person		
Précision globale	0,46	0,2	0,25	0,14	1	0,7
Rappel global	0,25	0,08	0,28	0,08	0,69	0,26
F-mesure globale	0,33	0,12	0,26	0,10	0,82	0,38

Trois types d'erreurs ont été remarquées : (1) erreur de type d'EN, (2) annotation partielle, (3) absence d'annotation. En particulier, la non reconnaissance de déclencheurs de lieux était à l'origine d'une partie des erreurs. Pour cette raison, l'adaptation au domaine a nécessité de mettre au point un dictionnaire de déclencheurs de lieux (boulevard, rue ...) afin d'améliorer la phase de réentraînement.

(1) Adaptation au domaine :

Adaptation 1 : La première adaptation consiste à effectuer un entraînement sur un extrait du roman « Le ventre de Paris » et une évaluation sur une autre partie du même roman. Les résultats de cette expérience nous montrent qu'un modèle entraîné et testé sur le même auteur, pour notre cas Zola, donne des résultats assez bons avec une précision globale de 1, un rappel global qui atteint 0,69 et une F-mesure entre de 0,82 (voir tab 1).

Ces résultats s'expliquent par le fait qu'un chapitre est un domaine de référence très restreint, dans lequel les mêmes noms de personne et de lieu ont tendance à se répéter. Donc entraîner sur une partie du chapitre produit de bons résultats d'annotation sur l'autre.

Adaptation 2 : C'est un entraînement sur un extrait du roman « Le ventre de Paris » (Zola) et une évaluation sur un roman d'un auteur différent, ici « César Birotteau » (Balzac).

Le même modèle d'apprentissage issu du roman de Zola (Adaptation 1) et testé sur celui de Balzac, donne des mesures assez faibles. Une précision de 0,7, un rappel de 0,26 et une F-mesure de 0,38. Ces résultats se justifient par le fait que le modèle est moins portable d'un auteur à l'autre.

(2) Progression :

Dans cette expérience le corpus d'entraînement composé d'extraits de Zola est progressivement augmenté afin de trouver la taille optimale de l'échantillon d'apprentissage. L'échantillon d'apprentissage est composé respectivement d'1/3 , 2/3 et 3/3 d'un chapitre et les modèles sont testés sur deux extraits différents de Zola et de Balzac.

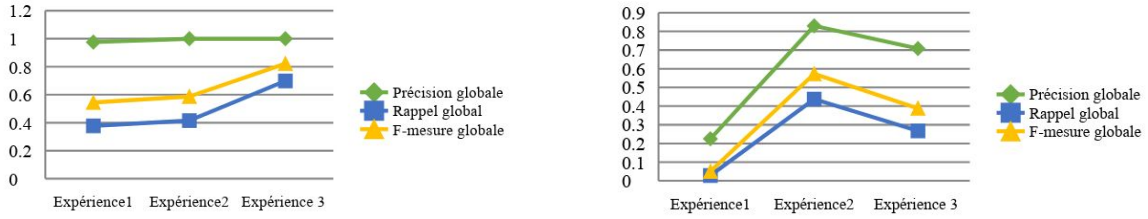

Figure 2. Progressions respectives pour Zola et Balzac

Pour le test sur le texte de Zola, en augmentant la taille du corpus d'entraînement la performance augmente progressivement. Cependant en effectuant un entraînement avec un texte de Zola et le testant sur le texte de Balzac, les résultats s'améliorent d'abord mais baissent quand le corpus d'entraînement est composé de 3 extraits. Ceci peut s'expliquer par le fait que le corpus d'entraînement est devenu trop adapté à Zola ce qui pourrait entraîner un problème de sur-apprentissage et manque de généralisation [Douglas M., 2004].

1.2. Adaptation et évaluation du système REDEN

REDEN [Frontini et al, 2015, Brando et al, 2016], est un outil de NEL fondé sur la théorie de graphes permettant la résolution d'EN s'appuyant sur des sources du Web de données. Cet outil prend en entrée un fichier XML/TEI tagué en entités nommées et produit en sortie le même fichier enrichi avec un identifiant, un IRI (acronyme pour

Internationalized Resource Identifier), pour chaque entité et un NIL pour les entités sans référent. Nous nous sommes intéressés uniquement aux lieux et avons adapté trois BC (DBpedia, Bnf, WikiData) à partir de requêtes SPARQL personnalisées. Chaque requête est relative à une seule base de connaissance et produit un dictionnaire qui est utilisé par REDEN pour rechercher les candidats. Nous avons enfin évalué REDEN à partir de textes annotés par SEM et pour chaque BC. Les métriques d'évaluation dépendent de la phase NEL concernée, à savoir (1) Recherche des candidats, (2) Sélection du bon candidat. Ces métriques étendent celles du REN (voir [Brando et al, 2016] pour les détails). Les résultats obtenus lors des expérimentations sont présentés dans le tableau 2.

Tableau 2. Résultats des expérimentations REDEN

Les mesures / BC	Candidate Precision	Candidate recall	NIL precision	NIL recall	Disambiguation accuracy	Taux D'ambigüité	Overall linking accuracy
DBpedia	1	0,816	0,367	1	none	0	0,834
BNF	0,760	0,630	0,580	0,972	1	0,005	0,7
Wikidata	0,912	0,830	0,440	1	1	0.29	0,85

Les métriques de la phase (1) nous permettent de déterminer son efficacité.
La « Candidate Precison », est de 1 pour la BC DBpedia, 0,912 pour Wikidata et 0,76 pour la Bnf. Donc en général, REDEN est capable de trouver l'IRI approprié dans la BC parmi un ensemble de candidats non vides.
En outre, pour le « Candidate recall » nous remarquons que les résultats avec DBpedia (0,816) et Wikidata (0,83) sont supérieurs aux résultats obtenus avec Bnf (0,63). Ceci se traduit par le fait que REDEN trouve plus de référents corrects avec DBpedia et Wikidata comparé à BnF. De plus, puisque nous nous intéressons aux toponymes dans la littérature issue du 19éme siècle, certains lieux ont changé de nom, par exemple « le fort de Bicêtre » existe dans la BNF sous le nom de « château de Bicêtre ». Aussi, les mentions sont des villes fictives comme « Plassans » qui existe dans DBpedia et Wikidata mais pas dans BNF.

Pour, évaluer la capacité de l'algorithme à produire des annotations NIL correctes pour les mentions n'ayant pas de réfèrent dans le *gold*. Nous observons que le résultat du « NIL Precision » est légèrement meilleur avec la base BNF 0,580 comparé à Wikidata 0,44 et DBpedia 0,367. Cependant, les résultats restent faibles, ceci s'explique par le fait que REDEN traite la phase de recherche des candidats avec l'algorithme des mesures de correspondance parfaites entre chaînes de caractères (*exact string match*).
D'autre part, le « NIL Recall » est élevé pour les trois BC, 1 pour DBpedia et Wikidata, 0,972 pour BnF. Ce qui se traduit par le fait, que comparé au NIL dans le *gold*, REDEN retourne des NILs corrects.
Quant aux mesures de la phase (2), le résultat obtenu pour la « Désambiguïsation accuracy » pour DBpedia est vide, comparé à la Bnf et Wikidata 1. Il est important de noter que cette mesure est intéressante quand on a des ensembles de candidats de taille supérieure à 1. Le taux d'ambiguïté est nul pour DBpedia, de 0,05 pour la Bnf et de 0,29 pour Wikidata. Ce qui signifie que la moyenne des ensembles de candidats ayant plus de 2 candidats est faible pour DBpedia ainsi que Bnf et légèrement plus importante avec Wikidata.
Enfin, la mesure de « Overall Linking » obtenue est de 0,834 pour DBpedia, 0,85 pour Wikidata et 0,7 pour BnF. REDEN a donc été efficace. Cette mesure essaie d'évaluer l'efficacité globale du système, et non par phase, et exprime la fiabilité de REDEN pour la tâche de résolution des entités nommées. Nous pouvons ainsi conclure que les trois bases sont efficaces pour la tâche de NEL, et donnent de résultats corrects. Cependant, Wikidata est légèrement meilleur que DBpedia et BnF.

Conclusion

Inspirés par le projet *Venice Time Machine,* développé à l'Ecole Polytechnique Fédérale de Lausanne par [Di Leonardo., 2015], notre travail s'avérera utile pour la création d'un volet littéraire d'un *Paris Time Machine*. En effet, ces projets ambitieux visent à la reconstruction du passé des grandes villes d'Europe à travers l'extraction d'informations à partir des documents historiques, y compris littéraires. Notre travail est très proche des travaux de [Boeglin et al., 2016], qui proposent une méthode pour localiser, cartographier et analyser les occurrences de lieux citées dans un corpus de 31 romans du 19ème siècle dont l'action se situe pour tout ou partie à Paris. Notre contribution est particulièrement d'avoir produit un modèle pour la reconnaissance d'EN dans les textes littéraires français ainsi que des BC du Web de données pertinentes pour la tâche NEL dans ce même contexte.

Afin d'illustrer nos propos, les figures 3 et 4, proposent deux rendus cartographiques possibles à partir de la chaîne de traitement proposée. Etant donné que les BC utilisés répertorient les coordonnées de localisation pour chaque entité de type lieu, il est donc possible, après la phase NEL, de rapatrier cette information (et d'autres) automatiquement par le biais des IRI. La première vue montre les lieux qui ont été mentionnés dans les romans et les marque avec un indicateur sur la carte. La deuxième vue nous permet de visualiser sur une carte pour chaque mention de lieu repérée dans le texte le nombre d'occurrences de cette EN dans les romans.

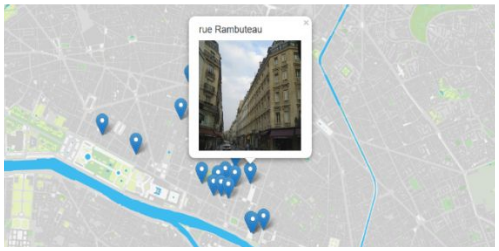

Figure 3. Vue numéro 1 de la cartographie

Figure 4. Vue numéro 2 de la cartographie

Références

Boeglin et al., 2016 : Boeglin N., Depeyre M., Joliveau T., Le Lay.F . Pour une cartographie romanesque de Paris au XIXe siècle. Proposition méthodologique. Conférence Spatial Analysis and GEOmatics, Actes de la conférence SAGEO'2016 - Spatial Analysis and GEOmatics Dec 2016, Nice, France, 2016.

Brando et al., 2015 : Brando.C, Frontini F., Ganascia J.: "Linked Data for toponym linking in French Literary texts", in Proceedings of the 9th Workshop on Geographic Information Retrieval; 2015.

Brando et al., 2016 : Brando C., Abadie N., Frontini F. : « Evaluation de la qualité des sources du Web de Données pour la résolution d'entités nommées ». Revue Ingénierie des Systèmes d'Informations, 2016.

Di Leonardo et al., 2015 : Di Lenardo, Isabella, Kaplan, Frédéric. Venice Time Machine : Recreating the density of the past, Digital Humanities 2015, Sydney, June 29 - July 3, 2015.

Douglas et al., 2014 : Douglas M. Hawkins,The Problem of Overfitting,School of Statistics, University of Minnesota, Minneapolis, Minnesota, 2014.

Dupont 2017 : Dupont Yoann. Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique. RECITAL, 2017, p. 42.

Frontini et al., 2015 : Francesca F., Brando C., and Ganascia J. : "Domain adapted named-entity linker using Linked Data » ; in Workshop on NLP Applications : Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems ;2015.

Lecluze et al., 2014 : Lecluze C., et Lejeune G., « DEFT2014, analyse automatique de textes littéraires et scientifiques en langue française » 21ème Traitement Automatique des Langues Naturelles, Marseille, 2014.

Raymond et al., 2010 : Raymond C., et Fayolle J., . Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. Dans Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10), Montréal, Canada, 2010.

Sagot et al., 2012 : Sagot B., Richard M., Stern R. . Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. Antoniadis G., Blanchon H., Sérasset G., . Traitement Automatique des Langues Naturelles (TALN), Jun 2012, Grenoble, France. 2 - TALN, 2012, Actes de la conférence conjointe JEP-TALN-RECITAL 2012. <hal-00703108>